Linear Regression

Jose Luis Contreras Santos, Antonio Javier González Ferrer, Alejandro González Pérez

enero, 2017

Abstract

An attempt to use linear regression to predict wine quality.

Introduction

The aim of this section is to use linear regression to model and predict a wine's quality based on its physicochemical attributes. To do this, we will consider the variable `quality` as a continuous variable ranging from 0 to 10. It is worth noting, however, that this variable is originally a categorical one which can only take one of the 11 integer values comprised between 0 and 10. Therefore, as predictions will be continuous, many of them will most likely be slightly off due to this discrepancy. This should not pose a problem, as the usual error metrics, such as RMSE, can be obtained anyways.

Model fitting

After splitting the data in training and test sets, the training set was used to fit a linear regression model. At first, all variables were used as explanatory variables for the model.

```
# Train and test dataset, split 80% (data has been shuffled previously, no need to sam
split = nrow(df)*0.8
train = df[1:split,]
test = df[split:nrow(df),]

# Fit the model using all variables
model1 = lm(quality~fixed_acidity+volatile_acidity+citic_acid
            +residual_sugar+chlorides+free_sulfur_dioxide
            +total_sulfur_dioxide+density+pH+sulphates+alcohol, data=train)
summary(model1)
```

```
##
## Call:
## lm(formula = quality ~ fixed_acidity + volatile_acidity + citic_acid +
##     residual_sugar + chlorides + free_sulfur_dioxide + total_sulfur_dioxide +
##     density + pH + sulphates + alcohol, data = train)
```

```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.8027 -0.4544 -0.0414  0.4593  2.7242
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)           5.556e+01  1.399e+01   3.973 7.21e-05 ***
## fixed_acidity         6.376e-02  1.793e-02   3.556 0.000380 ***
## volatile_acidity     -1.304e+00  8.613e-02 -15.136  < 2e-16 ***
## citic_acid           -1.199e-01  8.904e-02  -1.347 0.178173
## residual_sugar        4.162e-02  5.878e-03   7.080 1.64e-12 ***
## chlorides            -4.269e-01  3.602e-01  -1.185 0.236015
## free_sulfur_dioxide   6.192e-03  8.400e-04   7.371 1.96e-13 ***
## total_sulfur_dioxide -2.547e-03  3.093e-04  -8.236 2.23e-16 ***
## density              -5.481e+01  1.427e+01  -3.842 0.000124 ***
## pH                    4.673e-01  1.028e-01   4.546 5.59e-06 ***
## sulphates             7.452e-01  8.423e-02   8.847  < 2e-16 ***
## alcohol               2.663e-01  1.957e-02  13.609  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7338 on 5184 degrees of freedom
## Multiple R-squared:  0.2918, Adjusted R-squared:  0.2903
## F-statistic: 194.1 on 11 and 5184 DF,  p-value: < 2.2e-16
```

However, as the summary above indicates, the p-values for citic acid and chlorides are not low enough to reject the null hypothesis. Thus, we can consider that the contribution of these to the variance of quality is not significantly greater than 0, and so they have been removed from the final model. The resulting linear regression model is described below.

```
# Citic_acid removed from the model
model1 = update(model1, ~.-citic_acid)
# Check if chlorides can also be removed
summary(model1)
```

```
##
## Call:
## lm(formula = quality ~ fixed_acidity + volatile_acidity + residual_sugar +
##     chlorides + free_sulfur_dioxide + total_sulfur_dioxide +
##     density + pH + sulphates + alcohol, data = train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.7891 -0.4580 -0.0373  0.4598  2.7323
##
```

```
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)           5.554e+01  1.399e+01   3.971 7.26e-05 ***
## fixed_acidity         5.809e-02  1.743e-02   3.333 0.000866 ***
## volatile_acidity     -1.259e+00  7.940e-02 -15.852  < 2e-16 ***
## residual_sugar        4.136e-02  5.876e-03   7.038 2.20e-12 ***
## chlorides            -5.103e-01  3.549e-01  -1.438 0.150475
## free_sulfur_dioxide   6.232e-03  8.395e-04   7.424 1.33e-13 ***
## total_sulfur_dioxide -2.614e-03  3.053e-04  -8.563  < 2e-16 ***
## density              -5.479e+01  1.427e+01  -3.840 0.000124 ***
## pH                    4.770e-01  1.025e-01   4.652 3.38e-06 ***
## sulphates             7.409e-01  8.417e-02   8.801  < 2e-16 ***
## alcohol               2.642e-01  1.951e-02  13.544  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7339 on 5185 degrees of freedom
## Multiple R-squared:  0.2915, Adjusted R-squared:  0.2901
## F-statistic: 213.3 on 10 and 5185 DF,  p-value: < 2.2e-16
```

```r
# It can indeed be removed, do it
model1 = update(model1, ~.-chlorides)
summary(model1)
```

```
##
## Call:
## lm(formula = quality ~ fixed_acidity + volatile_acidity + residual_sugar +
##     free_sulfur_dioxide + total_sulfur_dioxide + density + pH +
##     sulphates + alcohol, data = train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.7877 -0.4560 -0.0367  0.4578  2.7385
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)           5.958e+01  1.370e+01   4.348 1.40e-05 ***
## fixed_acidity         6.160e-02  1.726e-02   3.569 0.000362 ***
## volatile_acidity     -1.275e+00  7.863e-02 -16.211  < 2e-16 ***
## residual_sugar        4.318e-02  5.738e-03   7.526 6.15e-14 ***
## free_sulfur_dioxide   6.185e-03  8.390e-04   7.372 1.94e-13 ***
## total_sulfur_dioxide -2.576e-03  3.042e-04  -8.469  < 2e-16 ***
## density              -5.899e+01  1.397e+01  -4.224 2.44e-05 ***
## pH                    5.068e-01  1.004e-01   5.046 4.67e-07 ***
## sulphates             7.161e-01  8.240e-02   8.690  < 2e-16 ***
```

```
## alcohol               2.637e-01  1.951e-02  13.517  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7339 on 5186 degrees of freedom
## Multiple R-squared:  0.2912, Adjusted R-squared:   0.29
## F-statistic: 236.8 on 9 and 5186 DF,  p-value: < 2.2e-16
```

According to the obtained model, the two variables with the highest influence on quality are `volatile acidity` and `density`, specially the latter. This can be misleading however, as it represents the variation in quality per unit of the input, and the scales of the input variables differ in their order of magnitude. Checking the distribution of *density*, for example, the difference between the maximum and minimun values is lower than 0.03, whereas $total_sulfur_dioxide$ has a range of over 400. In any case, some conclusions can be extracted from this output. For example, denser wines tend to have a smaller perceived quality due to the negative coefficient, and, on the contrary, those with higher alcohol contents obtain highger quality perception.

The adjusted R-squared coefficient of the resulting model is rather low, with a value of only 0.29. The regression overall p-value, however, is low enough for us to be confident that there exists a relationship between at least part of the input and the output variable, *quality*, performing better than just the simple constant model.

Assessing the model: assumptions

Let us further examine the quality of the linear regression model by checking if the assumptions are met. In particular, we are interested in testing if the residuals are independent, normal and have constant variance.

```
# A: Check the assumptions
# Linearity
raintest(model1) # Yes, linear
```

```
##
##  Rainbow test
##
## data:  model1
## Rain = 1.0067, df1 = 2598, df2 = 2588, p-value = 0.4325
```

```
# Test normality
jarque.bera.test(residuals(model1)) # Answer: No normality
```

```
##
##  Jarque Bera Test
##
## data:  residuals(model1)
```

```
## X-squared = 287.77, df = 2, p-value < 2.2e-16
```

```
# Equal variances
bptest(model1) # No constant variance
```

```
##
##  studentized Breusch-Pagan test
##
## data:  model1
## BP = 79.793, df = 9, p-value = 1.777e-13
```

```
# Testing independence of the residuals
dwtest(model1, alternative="two.sided") # Not independent
```

```
##
##  Durbin-Watson test
##
## data:  model1
## DW = 2.0093, p-value = 0.7377
## alternative hypothesis: true autocorrelation is not 0
```

The conducted tests have different results. First of all, the Rainbow test tests the linear relationship between the response and the linear predictor. The p-value is high enough to not reject the null hypothesis (0.4325) but in other datasets this value is close to 1. On the other hand, the Durbin-Watson test returns a high p-value (0.7377), sign that there is not enough evidence to reject the null hypothesis of autocorrelation of the residuals. Hence, the residuals can be considered as independent. The results for the Jarque-Bera and Breusch-Pagan tests, however, are not so positive. According to their results, residuals are neither normal nor homoscedastic.

Although the residual plots are unusual due to the clustering around the integer values, they confirm the previous results. The Normal Q-Q plot is in line with the results of the Jarque-Bera test, showing departures from normality, specially in the first quartiles. From the Cook's distance plot, a significant outlier can be seen, corresponding to observation 4446. Even if we remove this outlier, the statistical assumptions remain the same.

Evaluation of results

The final step is to evaluate the performance of the model. In order to do so, let us measure the accuracy of predictions by calculating the value of our error metric of choice: RMSE. The standard deviation of our data will be used as a reference to assess the correctness of RMSE values.

```
# Standard deviation of our data
sd(df$quality)
```

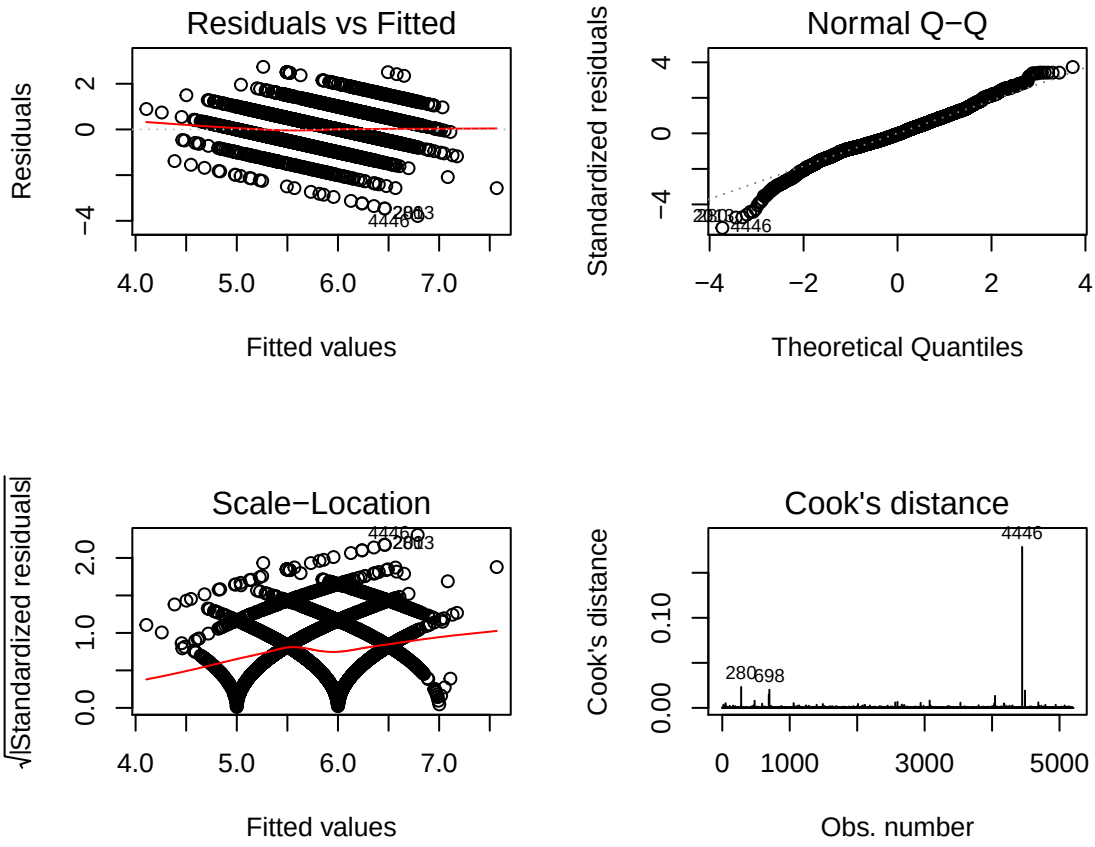

Figure 1: Residual plot

```
## [1] 0.8732716
```

```
# 2. Assess the model. B: Evaluate performance
predicted <- predict(model1, test)
rmse <- (sqrt(mean((test$quality - predicted)^2)))
rmse
```

```
## [1] 0.7426711
```

As the results above show, our predicted variable *quality* has a standard deviation of $\sigma = 0.87$. When fed with the test set, the linear regression model outputs predictions with an $RMSE = 0.74$. Therefore, and in line with what has been stated before, the interpretation of these results is positive, as RMSE is lower than the data's standard deviation.

Alternative model based on the sweetness of wines

We can try to improve the validation and interpretation of the model by defininig a new categorical variable, $residual_sugar2$, which divides wines in three groups according on their sweetness[1]. The sweetness of the wines is defined by its residual sugar, and commonly the classification is `dry` wines for values up to 4 g/l, `medium_dry` up to 12 g/l, `medium` up to 45 g/l and `sweet` more than 45 g/l. In out dataset, the highest value for `residual_sugar` is 18, thus we only consider the first three clusters.

```
# Adding the new attribute
df$residual_sugar2 <- df$residual_sugar
df$residual_sugar2[df$residual_sugar < 4.0] <- "dry"
df$residual_sugar2[df$residual_sugar >= 4.0 & df$residual_sugar < 12.0] <- "medium dry"
df$residual_sugar2[df$residual_sugar >= 12.0] <- "medium"

df$residual_sugar2 = as.factor(df$residual_sugar2)
df$residual_sugar2=relevel(df$residual_sugar2, ref="dry")

# Update train and test sets
train = df[1:split,]
test = df[split:nrow(df),]

model2 = update(model1, ~.+residual_sugar2+residual_sugar:residual_sugar2)
summary(model2)
```

```
##
## Call:
## lm(formula = quality ~ fixed_acidity + volatile_acidity + residual_sugar +
##     free_sulfur_dioxide + total_sulfur_dioxide + density + pH +
##     sulphates + alcohol + residual_sugar2 + residual_sugar:residual_sugar2,
##     data = train)
##
```

```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.8172 -0.4577 -0.0354  0.4526  2.6725
##
## Coefficients:
##                                              Estimate Std. Error t value
## (Intercept)                                 6.881e+01  1.387e+01   4.962
## fixed_acidity                               6.448e-02  1.725e-02   3.738
## volatile_acidity                           -1.311e+00  7.870e-02 -16.663
## residual_sugar                              1.519e-01  2.304e-02   6.591
## free_sulfur_dioxide                         5.922e-03  8.407e-04   7.044
## total_sulfur_dioxide                       -2.612e-03  3.065e-04  -8.523
## density                                    -6.835e+01  1.413e+01  -4.837
## pH                                          5.272e-01  1.004e-01   5.253
## sulphates                                   7.053e-01  8.240e-02   8.560
## alcohol                                     2.448e-01  1.998e-02  12.249
## residual_sugar2medium                       8.122e-01  1.818e-01   4.467
## residual_sugar2medium dry                   3.532e-01  7.809e-02   4.523
## residual_sugar:residual_sugar2medium       -1.451e-01  2.459e-02  -5.900
## residual_sugar:residual_sugar2medium dry   -1.265e-01  2.349e-02  -5.387
##                                            Pr(>|t|)
## (Intercept)                                7.20e-07 ***
## fixed_acidity                              0.000187 ***
## volatile_acidity                            < 2e-16 ***
## residual_sugar                             4.80e-11 ***
## free_sulfur_dioxide                        2.11e-12 ***
## total_sulfur_dioxide                        < 2e-16 ***
## density                                    1.36e-06 ***
## pH                                         1.55e-07 ***
## sulphates                                   < 2e-16 ***
## alcohol                                     < 2e-16 ***
## residual_sugar2medium                      8.11e-06 ***
## residual_sugar2medium dry                  6.24e-06 ***
## residual_sugar:residual_sugar2medium       3.87e-09 ***
## residual_sugar:residual_sugar2medium dry   7.49e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7312 on 5182 degrees of freedom
## Multiple R-squared:  0.2969, Adjusted R-squared:  0.2952
## F-statistic: 168.4 on 13 and 5182 DF,  p-value: < 2.2e-16
```

```
# Test linearity
raintest(model2) # Yes, linear
```

```
##
##  Rainbow test
##
## data:  model2
## Rain = 1.0076, df1 = 2598, df2 = 2584, p-value = 0.4236
```

```
# Test normality
jarque.bera.test(residuals(model2)) # Answer: No normality
```

```
##
##  Jarque Bera Test
##
## data:  residuals(model2)
## X-squared = 289.9, df = 2, p-value < 2.2e-16
```

```
# Equal variances
bptest(model2) # No constant variance
```

```
##
##  studentized Breusch-Pagan test
##
## data:  model2
## BP = 93.627, df = 13, p-value = 2.818e-14
```

```
# Testing independence of the residuals
dwtest(model2, alternative="two.sided") # Not independent
```

```
##
##  Durbin-Watson test
##
## data:  model2
## DW = 2.0065, p-value = 0.8158
## alternative hypothesis: true autocorrelation is not 0
```

```
# 2. Assess the model. B: Evaluate performance
predicted2 <- predict(model2, test)
rmse <- (sqrt(mean((test$quality - predicted2)^2)))
rmse
```

```
## [1] 0.74352
```

In comparison with the previous model, the R-squared coeffient has improved slightly. The model's assumptions, however, remain the same: residuals are linear, but they are not normal and their variance is still not constant either. RMSE does not change significantly, differing in less than 0.01 from the previous measure.

Overall, the improvement in R-squared does not justify the effort to include a new attribute, as a) it is very slight, b) the model's assumptions are still not met, and c) RMSE has not improved as a consequence of it. It is true, however, that by looking at the coefficients, some

information can be derived about the impact of sweetness on quality. Specifically, medium wines seem to have higher quality rankings than medium-dry ones, which in turn are ranked better than dry wines.

Conclusions

According to evaluation results, the model has a rather good performance at predicting wine quality. Even though the it does not have a very high R-squared coefficient, the RMSE metric is low enough for us to consider it an interesting option to predict a wine's quality. Some remarks can be made, however. First of all, our input data is not uniformly distributed, with a majority of the observations taking quality values between 4 and 6. Thus, our model could be slightly overfitted for those kind of values, and there is a risk that the precision of results could drop if the input data happened to contain more extreme values.

Besides, it would be interesting to try a different approach to this problem, perhaps using a classification technique for prediction instead, although linear regression seems better suited given the features of the output variable.

Bibliography

[1] Terms used to indicate sweetness of wine. https://en.wikipedia.org/wiki/Sweetness_of_wine#Residual_sugar